COMP00162 – Report

Title: Exploration of volatility and potential solutions to predictions

Abstract:

Introduction (2 pages max):

* Background
* Problem to be solved: stock prediction
* The stock market can be an incredibly unpredictable ecosystem encompassing uncertainty at various aspects of the domain.
* This volatility can be attributed to various economic factors and market forces which have an impact on the demand, cost and accessibility of stocks. Other factors include rates associated with inflation, interest tax changes and changes within corporate industries.
* The process of investing in commodities is not easy and hence the before mentioned conditions make it difficult to predict future rates of stocks.
* In the past 5 years, there has been significant research into the use of machine learning models for predicting prices of single assets as well as whole portfolios. The next question that might be asked is what makes machine learning a suitable tool for forecasting stocks?
* Machine learning focusses on acquiring implicit information from the data, extracting the meaning in patterns and using it to improve on predictions through a learning process. Furthermore, emphasis should be highlighted on ‘learning’ as an increase in data volume significantly increases this process. Furthermore, the vast availability of stock data is effective with many data hungry ML algorithms such as neural networks.
* [Stock Market Prediction Using LSTM Recurrent Neural Network - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S1877050920304865)
* As shown in the paper by S. Makridakis, machine learning within industry still requires further evidence that it could be used a standard instrument for making projections.
* <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889>
* One of the most common machine learning algorithms is the Random Forest model. The random forest can be considered part of a family of traditional machine learning algorithms, can be used for regressive as well as classification tasks.
* The processes within the methodology of these algorithms are significantly different to that used in deep learning algorithms. A key difference is that deep learning utilises neural networks, an architecture consisting of input layers, hidden layers and output layers. Neural networks can be considered a form of ‘black box’ systems because analysis of their structure does not provide any information on the structure of the function it tries to approximate.
* This means there is no transparency in the transformation which produces the outputs.
* <https://www.frontiersin.org/articles/10.3389/fpsyt.2020.551299/full>
* Additionally, Recurrent Neural Networks (RNNs) are a group of neural networks which incorporate the dependency of the output and any one input on the previous inputs. This ‘memory’ property allows for the network to process sequential data such as the stock time-series.
* Specifically, Long-Short-Term-Memory networks are RNNs that utilise a special recurrent unit (module) that are capable of learning long-term dependencies as well short-term ones within sequences. This is functionally important in the context of stock time-series because current prices of commodities might be dependent on prices and events that have happened relatively long time ago. Furthermore, it allows for the model to make more accurate long-term forecasts into the future.
* <https://www.ijert.org/a-review-on-using-long-short-term-memory-for-prediction-of-stock-price>
* <https://www.springerprofessional.de/en/learning-accurate-lstm-models-of-business-processes/17094226>
* Various features of the LSTM make it more suitable than a simple RNN. Firstly, the presence of the cell state plays a key role in one module of an LSTM. The cell state stores the long-term dependencies which allow the present unit to receive information from inputs at time steps that occurred significantly back in time. Additionally, the derivatives of the cell state lessen the effects of vanishing and exploding gradients.

<https://hcis-journal.springeropen.com/articles/10.1186/s13673-020-00242-w>

* The vanishing gradient can be significant in simpler RNNs because the recurrent weight within the temporal loops will tend to 0 on continuous multiplication during the updates.
* LSTMs aim to reduce this by setting the recurrent weight to 1, given that greater values lead to exploding gradients and smaller values lead to vanishing gradients.

<https://direct.mit.edu/neco/article-abstract/9/8/1735/6109/Long-Short-Term-Memory?redirectedFrom=fulltext>

* The following investigation aims to further delve into volatility and how LSTM recurrent units can be used as time series models for the prediction of stock net returns. Furthermore, the performance of the LSTM model will be compared with those of regression random forest as a baseline model.
* Discrete time series model for stocks predictions
* Provocative question (why is the problem not solvable by old methods? What attributes of the lstms allow for a better solution?)
* Literature review (10 papers)
* ARIMA models
* What is it
* Why has it been used in the past

<https://www.sciencedirect.com/science/article/pii/S2405918818300060>

* Support vector machines (regression)
* What is it
* Why has it been used so far

<https://www.sciencedirect.com/science/article/pii/S1877050920304865>

* Recurrent neural network are

<file:///C:/Users/micha/Downloads/jrfm-13-00181.pdf>

* Random forest with boosting and bagging

[Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns by Benjamin Moritz, Tom Zimmermann :: SSRN](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740751)

* Why a machine learning model is suitable for the problem
* Current models used such as ARIMA, support vector machines, single and multi-time series models,

Methodology:

* Mathematical foundation of the model implemented (formulas and equations) (baseline model and the complex model) (what hardware has been used to run the code and produce the results)
* Description and initial analysis of the time series data
* Logic and functioning of the experiment
* Main methodology
* Brief description of the data, were it was obtained, description of the exploratory data analysis, describe and explain the data cleaning

Results:

* Detailed presentation and analysis of the results
* Figures and tables showing the key findings in the results
* Comparison of the main machine learning methodology with a baseline model

Discussion:

* Interpretation of the results and connected with the main assumptions of the methodology
* Outline of the advances
* Perspectives for the future
* Improvements and limitations

Conclusion:

Bibliography:

Proposes

Notes:

* Overfitting and underfitting
* Exploding and vanishing gradients
* Problems with normal recurrent neural networks